

# A Bayesian Approach to Estimation of Speaker Normalization Parameters

Dhananjay Ram<sup>a,b,\*</sup>, Debasis Kundu<sup>c</sup>, Rajesh M. Hegde<sup>c</sup>

*dhananjay.ram@idiap.ch*

*{kundu,rhegde}@iitk.ac.in*

<sup>a</sup>*Idiap Research Institute, Martigny, Switzerland*

<sup>b</sup>*Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

<sup>c</sup>*Indian Institute of Technology Kanpur, India*

---

## Abstract

In this work, a Bayesian approach to speaker normalization is proposed to compensate for the degradation in performance of a speaker independent speech recognition system. The speaker normalization method proposed herein uses the technique of vocal tract length normalization (VTLN). The VTLN parameters are estimated using a novel Bayesian approach which utilizes the Gibbs sampler, a special type of Markov Chain Monte Carlo method. Additionally the hyperparameters are estimated using maximum likelihood approach. This model is used assuming that human vocal tract can be modeled as a tube of uniform cross section. It captures the variation in length of the vocal tract of different speakers more effectively, than the linear model used in literature. The work has also investigated different methods like minimization of Mean Square Error (MSE) and Mean Absolute Error (MAE) for the estimation of VTLN parameters. Both single pass and two pass approaches are then used to build a VTLN based speech recognizer. Experimental results on recognition of vowels and Hindi phrases from a medium vocabulary indicate that the Bayesian method improves the performance by a considerable margin.

**Keywords:** Speaker Normalization, Bayesian Estimation, Vocal Tract Length Normalization (VTLN), Mean Square Error (MSE), Mean Absolute Error (MAE), Hyperparameters

---

## 1. Introduction

One of the biggest challenges in the design of an automatic speech recognizer (ASR) is to compensate for the speaker variability. It is caused by the acoustic variations introduced in the signals of same utterance spoken by different speakers. In spite of these acoustic variations, humans can recognize utterances of different speakers very easily. But, speech recognizer cannot recognize same words uttered by different speakers very well due to these variations (1). Different types of speaker normalization methods are used to enhance the recognition accuracy when different speakers are using same recognizer (2; 3; 4). Many researchers have approached the problem of normalization using only formants of vowels. Nordstrom and Lindblom (5) determined a constant scale factor depending on the ratio of third formant of subject to that of reference speaker. Later, Fant (6) argued that uniform scaling is a very simple approximation, so the scale factor should be dependent on both vowel category and formant number. Then, Miller (7) applied formant ratio theory to normalization problem, which claims that vowels are relative patterns. A detailed study of these vowel normalization procedures by Nicholas Flynn can be found in (8).

Warping functions are also used to reduce differences between the spectra of subject and reference speakers. Different types of warping functions have been used in past, namely Linear warping function (9), Piecewise Linear warping function (10), Affine warping function (11), Non-linear warping function (12) etc.

---

\*Corresponding author

Eide and Gish (13) have developed a warping function based on the median position of third formant in speech of the speaker under consideration. Stevens and Volkmann (14) have proposed a non-linear warping function based on their perceptual studies on speech signals. The well known bilinear transform is also used as a warping function in (15). A likelihood maximization technique of speaker normalization is proposed in (2). In this method, warping factor for a particular speaker is chosen by maximizing the likelihood of a hypothesis in an iterative manner at the output of a recognizer. In (10), Gaussian Mixture Model is used to represent a class of standard speakers. The classes are assigned warping factors from a set of values. Different speaker are assigned to one of these classes using likelihood maximization. Lee and Rose (9) have proposed a similar method. But, they have used maximization of likelihood of the Hidden Markov Model (HMM) for normalization as well as recognition. Researchers have also used normalization in the feature domain. Acero and Stern (15) have proposed an affine transformation of the cepstral features. Cox (16) has shown that vocal tract length normalization (VTLN) can be implemented with filter banks and directly applied this method in the feature domain. Later in (17), a linear transformation approach using Dynamic Frequency Warping (DFW) has been proposed for normalization.

In this paper an affine model based speaker normalization method is presented. The model parameters are estimated using Bayesian estimation as well as error function minimization technique. This work justifies the use of Bayesian method of parameter estimation, which is also supported by the experimental results. Techniques like bandwidth adjustment and frequency bin adjustment are discussed which are required for the application of speaker normalization in the speech recognizer. The rest of the paper is organized as follows. Section 2 describes the model and error function minimization technique is used to estimate the model parameters. Section 3 proposes Bayesian estimation method to estimate the affine model parameters. The need of hyperparameters estimation is also discussed in this Section and the hyperparameters are estimated using likelihood maximization. Section 4 presents Formant frequency based vowel recognizer and HMM based speech recognizer, and experimental condition to perform different types of experiments on these recognizers. The incorporation of normalization method with recognizer for training as well as testing on a database is described using block diagrams. Experimental results for both vowel as well as word recognition are shown in the same Section. Finally, conclusions are presented in Section 5.

## 2. Speaker Normalization using Frequency Warping

An affine model (18) for speaker normalization is introduced in this section. The model parameters,  $\alpha$  and  $\kappa$  are estimated using principle of Error Function Minimization. However it is found that, for a range of values of  $\alpha$ , estimated value of  $\kappa$  is not reliable. An adjustment technique is suggested for estimated values of parameters to get more reliable estimates.

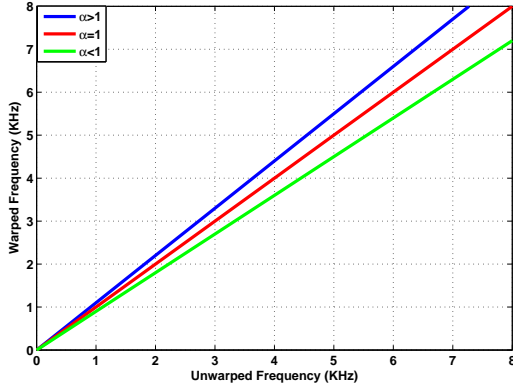
### 2.1. Affine Model for Vocal Tract Length Normalization

The affine model used for speaker normalization in an earlier work (18) is given by

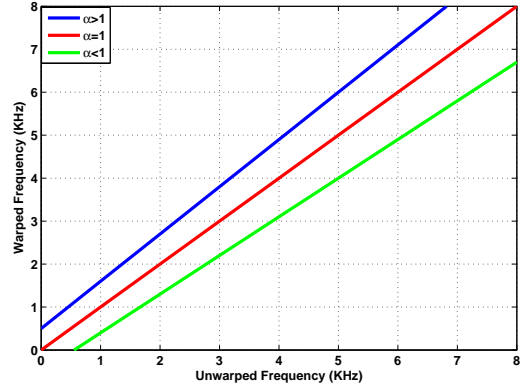
$$\mathbf{Y} = \alpha \mathbf{X} + \kappa(\alpha - 1)\mathbf{1} \quad (1)$$

where, the vectors,  $\mathbf{Y}$  and  $\mathbf{X}$  represent the formant frequency vector for reference and subject speakers respectively. Formant frequency vector is constructed by concatenating formants of all utterances corresponding to a particular speaker. So, length of this vector depends on the database under consideration.  $\alpha$  and  $\kappa$  are respectively, the speaker dependent and independent parameters.  $\mathbf{1}$  is a vector of 1s, i.e.  $\mathbf{1} = [1 \ 1 \dots 1]^T$ . The dimension of  $\mathbf{1}$  is same as the dimension of  $\mathbf{Y}$  or  $\mathbf{X}$ .

It can be easily seen from Figure 1, that the affine model has a shift factor in addition to the scaling factor used in the popular linear model for VTLN. The linear warping function always passes through the origin, whereas the affine warp function passes through the origin only for  $\alpha = 1$ . In other cases, it has a positive intercept for  $\alpha > 1$  and negative intercept for  $\alpha < 1$ . The warped spectrum gets shifted due to non-zero intercept of the normalization function. The shift is towards right or left for values of  $\alpha$  greater than 1 or less than 1 respectively.



(a) Linear Model



(b) Affine Model

Figure 1: Comparison of warping functions for different values of  $\alpha$  and constant  $\kappa$

The shift factor in Equation (1) is not same for different subject speakers. But, the number of parameters has not been doubled to achieve this, compared to the number of parameters in the linear model. It is achieved simply by making the shift factor, a function of the speaker dependent scaling factor. In this way, the increase in the number of parameters is only one, but an effect of as many parameters as the number of speakers has been achieved. Thus, a lesser accurate model is obtained compared to the scenario when the shift factors are also speaker dependent. It is implemented to keep a check on the number of parameters to be estimated, which effectively reduces computational complexity of the system. This same affine model was used in (18). But, the authors have used it to come up with a universal warping function having the same parametric form as the mel scale.

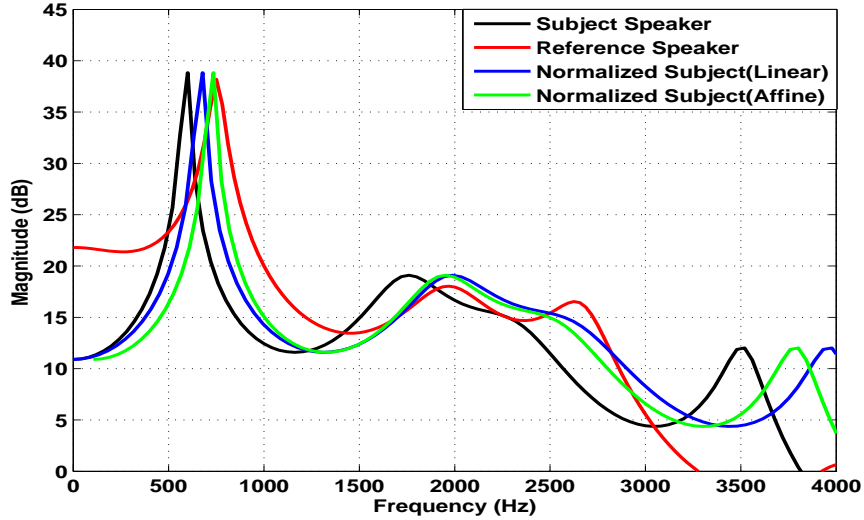


Figure 2: Normalization of spectrum of the vowel 'ae'

Figure 2 shows four different LPC smoothed spectra of same vowel, 'ae'. In this figure, two spectra represent utterances by a subject speaker and a reference speaker, whereas the other two spectra are normalized spectra of the subject speaker to make it closer to the reference speaker in terms of formant peaks. The two normalized spectra are obtained using popularly used linear model and an affine model respectively.

The first and second formant frequency values are given in Table 1. In this table, values in first row shows formants of reference speaker. The following rows indicate formant of normalized spectra of subject speaker using linear and affine model, respectively.

Table 1: Table Showing First Two Formant Frequencies of Spectra Presented in Figure 2

| Normalization<br>Function | Formant Frequency  |                     |
|---------------------------|--------------------|---------------------|
|                           | First Formant (Hz) | Second Formant (Hz) |
| Reference                 | 750                | 1970                |
| Subject                   | 600                | 1760                |
| Linear Function           | 678                | 2000                |
| Affine Function           | 735                | 1960                |

The formant frequency values of normalized spectra clearly show that, affine model gives a better match compared to the linear model, in terms of formant frequencies.

### 2.2. Affine Model Parameter Estimation using Error Function Minimization Technique

An error function minimization technique is used for estimating the parameters  $\alpha$  and  $\kappa$  of the affine model mentioned earlier. The model can be written as

$$\mathbf{Y} = \alpha \mathbf{X} + \kappa(\alpha - 1)\mathbf{1} + \boldsymbol{\epsilon} \quad (2)$$

where,  $\boldsymbol{\epsilon}$  is the error vector. The estimated formant frequency vector after applying these parameters on  $\mathbf{X}$  is

$$\hat{\mathbf{Y}} = \alpha \mathbf{X} + \kappa(\alpha - 1)\mathbf{1} \quad (3)$$

The parameter  $\kappa$  is assumed to be speaker dependent. Then, the estimated formant frequency vector can be written as

$$\hat{\mathbf{Y}}_{ij} = \alpha_{ij} \mathbf{X}_j + \kappa_{ij}(\alpha_{ij} - 1)\mathbf{1} \quad (4)$$

where  $\alpha_{ij}$  and  $\kappa_{ij}$  are the parameters for the  $j$ -th subject speaker with respect to  $i$ -th reference speaker.

The above Equation (4), when averaged over all the reference speakers gives the following

$$\hat{\mathbf{Y}}_j = \alpha_j \mathbf{X}_j + \kappa_j(\alpha_j - 1)\mathbf{1} \quad (5)$$

Now, the parameters,  $\kappa_{ij}$  and  $\alpha_{ij}$  in Equation (4) are estimated by considering every possible pair of subject and reference speaker and minimizing an error function. Mean Square Error (MSE) and Mean Absolute Error (MAE) have been used as error functions.

In MSE, the error function to be minimized is

$$\begin{aligned} \boldsymbol{\epsilon}_{ij} &= (\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij})^T (\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij}) \\ &= (\mathbf{Y}_{ij} - \alpha_{ij} \mathbf{X}_j + \kappa_{ij}(\alpha_{ij} - 1)\mathbf{1})^T (\mathbf{Y}_{ij} - \alpha_{ij} \mathbf{X}_j + \kappa_{ij}(\alpha_{ij} - 1)\mathbf{1}) \end{aligned} \quad (6)$$

In MAE, the error function to be minimized is

$$\begin{aligned} \boldsymbol{\epsilon}_{ij} &= |\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij}|^T \mathbf{1} \\ &= |\mathbf{Y}_{ij} - \alpha_{ij} \mathbf{X}_j + \kappa_{ij}(\alpha_{ij} - 1)\mathbf{1}|^T \mathbf{1} \end{aligned} \quad (7)$$

where, the operator  $|\cdot|$  gives element-wise absolute value of corresponding vector. Two different estimates of  $\kappa_{ij}$  and  $\alpha_{ij}$  are obtained by minimizing Equations (6) and (7). For each of the above mentioned functions, joint optimization has been used for estimating  $\kappa_{ij}$  and  $\alpha_{ij}$ . Nelder-Mead method (19) is implemented to minimize both Equations (6) and (7). This is a well defined numerical method for problem for which

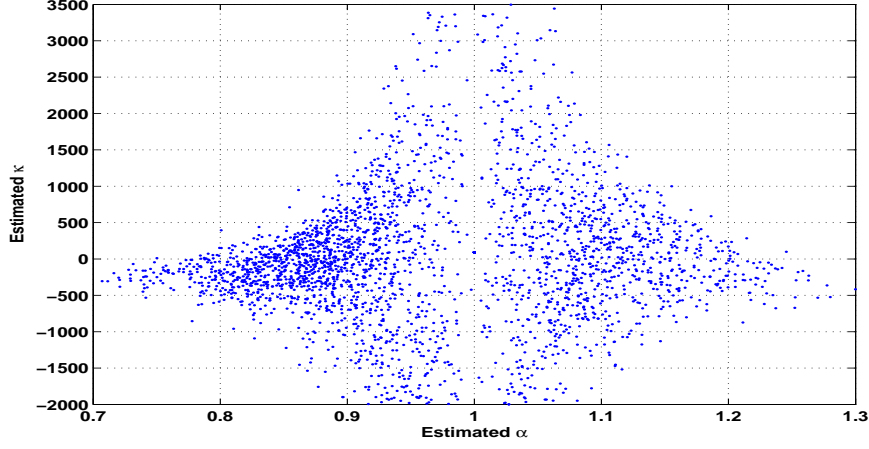


Figure 3: Illustration of large variability of  $\kappa$  near  $\alpha = 1$

derivatives may not be known. Now, it is assumed that, there are  $m$  subject speakers and  $n$  reference speakers. So, the average estimated formant frequency vector for the  $j$ -th subject speaker will be

$$\hat{\mathbf{Y}}_j = \left( \frac{1}{n} \sum_{i=1}^n \alpha_{ij} \right) \mathbf{X}_j + \left( \frac{1}{n} \sum_{i=1}^n \kappa_{ij} (\alpha_{ij} - 1) \right) \mathbf{1} \quad (8)$$

Now, comparison of Equations (5) and (8) gives the following,

$$\alpha_j = \frac{1}{n} \sum_{i=1}^n \alpha_{ij} \quad (9) \quad \kappa_j = \frac{\sum_{i=1}^n \kappa_{ij} (\alpha_{ij} - 1)}{\sum_{i=1}^n (\alpha_{ij} - 1)} \quad (10)$$

The shift factor,  $\kappa$  for the entire database is obtained by averaging  $\kappa_j$  over all the subject speakers

$$\kappa = \frac{1}{m} \sum_{j=1}^m \kappa_j \quad (11)$$

An ardent observation of Equation (1) indicates that, for  $\alpha = 1$ ,  $\kappa$  can take any value from  $-\infty$  to  $+\infty$ . Hence,  $\kappa$  cannot be estimated reliably. Also, the estimated values of  $\kappa$  show a great amount of variance for values of  $\alpha$  close to 1. Figure 3 shows the variation of estimated  $\kappa$  with corresponding estimates of  $\alpha$ . In other words, estimated values of  $\kappa$  are also not reliable for the values of  $\alpha$  near 1. An outlier adjustment techniques is applied to deal with this problem. An outlier is defined as an estimated value that deviates markedly from other estimated values of a parameter. In this method, a range of  $[0, L]$  is chosen for estimated  $\kappa_{ij}$ , for  $i$ -th reference speaker and  $j$ -th subject speaker. The following expression is used to adjust the values of  $\kappa_{ij}$  to the boundary.

$$\kappa_{ij} = \max(0, \min(\kappa_{ij}, L)) \quad (12)$$

The adjusted values for  $\kappa_{ij}$  are 0 and  $L$  for  $\kappa_{ij} < 0$  and  $\kappa_{ij} > L$  respectively. The outlier adjustment technique discussed here is an empirical method based on observations. Note that, the MMSE estimation and corresponding outlier adjustment technique has been proposed in (18), whereas the MMAE estimation technique is proposed in this paper. This technique has been successfully applied in (18) with affine model. The same method is applied here in order to compare its results with the proposed Bayesian method. Also, the results in Section 4 show that, this method improves performance of vowel recognizer for gender

dependent normalization in most cases but not for all.

Now, the algorithm to evaluate affine model parameters using error function minimization technique is given below.

---

**Algorithm 1** Estimation of Parameters  $\alpha$  and  $\kappa$  using MMSE and MMAE

---

- 1: **Formant Frequency Vectors** : Formant frequencies are extracted (20) from all utterances to construct formant frequency vector for subject speaker ( $\mathbf{X}$ ) and reference speaker ( $\mathbf{Y}$ ).
  - 2: **Error function** : MSE and MAE for  $i$ -th reference speaker and  $j$ -th subject are constructed as shown in Equations (6) and (7) respectively.
  - 3: **Estimation of  $\alpha_{ij}$  and  $\kappa_{ij}$**  :  $\alpha_{ij}$  and  $\kappa_{ij}$  are estimated by minimizing Equations (6) and (7) using Nelder-Mead method. These error functions give two different estimates of the same parameter.
  - 4: **Final  $\alpha$**  : Finally,  $\alpha_j$  for  $j$ -th subject speaker is obtained by averaging  $\alpha_{ij}$  over all reference speakers.
  - 5: **Outlier adjustment for  $\kappa_{ij}$**  : If  $\kappa_{ij}$  values are outside a given range  $[0, L]$ , its values are adjusted to this range using Equation (12).
  - 6: **Final  $\kappa$**  : First  $\kappa_j$  is calculated using Equation (10) and then, it is averaged over all subject speakers to obtain final  $\kappa$  for the database.
- 

MMSE and MMAE estimation techniques have been used in this section to estimate the affine model parameters. But, these techniques require  $\kappa$  to be speaker dependent, which is essentially opposite to the premise of the model, where it is assumed to be speaker independent. Also, the huge variance of  $\kappa$  is controlled using an outlier adjustment technique, in which the adjustment range may change depending on the database. In order to overcome these problems, Bayesian estimation method is proposed in the following section.

### 3. Bayesian Approach to Estimation of Affine Model Parameters

In this section, Bayesian Estimation has been introduced for estimating the parameters. First, the motivation behind using the Bayesian estimation is discussed. The motivation is followed by an elaborate mathematical description of Bayesian estimation technique for the problem at hand. Additionally, Gibbs sampler is used for numerical estimation of the model parameters. Subsequently, maximum likelihood estimation technique is used to estimate hyperparameters of the model.

#### 3.1. Motivation

The proposed Bayesian method for affine model parameter estimation is broadly motivated by the following observations,

- The speaker independent parameter  $\kappa$  of the affine model described in Section 2.1 is not estimable under the frequentist set-up, if true value of  $\alpha = 1$ . Even if  $\alpha \neq 1$  but very close to 1, which is usually the case in practice, the least squares estimate or the maximum likelihood estimate (under the assumption of Gaussian error) of  $\alpha$  becomes very unreliable. It has been observed in the simulation study presented in Figure 3 that the variance of the estimator of  $\kappa$  is very high in such a situation.
- On the other hand under the Bayesian framework, because of the random nature of  $\alpha$ ,  $\kappa$  is always estimable. Since in practice  $\alpha$  is very close to 1, if we can use our prior knowledge on  $\alpha$ , very reliable Bayes estimate of  $\kappa$  can be obtained. Even if we do not have any prior knowledge of  $\alpha$ , the data driven prior works much better than the usual least squares or maximum likelihood estimator.
- Also, there is no need to assume  $\kappa$  to be speaker dependent and average out over all speakers to obtain final estimate of  $\kappa$ , which is speaker independent. It can be seen later in this section that, the independence property of  $\kappa$  is preserved in the Bayesian framework and the large variance of  $\kappa$  is reduced in its final estimate.

Since, now a days modern Bayesian technique like MCMC (21) is available, in this case Bayesian inference seems to be the obvious choice.

### 3.2. Bayesian Estimation Technique

The model parameters are considered to be random variables for using Bayesian Estimation (22) method. These random variables are assumed to follow a distribution depending on prior knowledge about the parameter and hence these are called the prior distribution. Thereupon the posterior distributions for model parameters are derived using prior distributions and the observed data. The parameters corresponding to the prior distributions are called hyperparameters. The hyperparameters are estimated first in order to obtain better estimate for model parameters. Different kinds of estimates can be obtained using the posterior distribution depending the loss function under consideration.

The affine model for  $i$ -th reference speaker and a fixed subject speaker is given by

$$\mathbf{Y}_i = \alpha_i \mathbf{X} + \kappa(\alpha_i - 1)\mathbf{1} + \boldsymbol{\epsilon}_i \quad (13)$$

where,  $\boldsymbol{\epsilon}_i$  is the error vector. It is assumed to be a multivariate gaussian with zero mean i.e.  $\boldsymbol{\epsilon}_i \sim \mathcal{N}_r(\mathbf{0}, \sigma^2 \mathbf{I}_{r,r})$ . Here,  $r$  is the length of formant frequency vector of a speaker in the database,  $\mathbf{0}$  is a vector of zeros of length  $r$  and  $\mathbf{I}_{r,r}$  is an identity matrix of size  $(r \times r)$ . From the above Equation, the mean and variance of  $\mathbf{Y}_i$  can be calculated as

$$\mu_i = E(\mathbf{Y}_i) = \alpha_i \mathbf{X} + \kappa(\alpha_i - 1)\mathbf{1} \quad (14)$$

$$\Sigma_{\mathbf{Y}_i} = \sigma^2 \mathbf{I}_{r,r} \quad (15)$$

Using Equations (14) and (15) the probability density function of  $\mathbf{Y}_i$  can be written as

$$f_y(\mathbf{Y}_i | \kappa, \alpha_i, \sigma) = \frac{1}{(2\pi\sigma^2)^{r/2}} e^{\frac{-(\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}} \quad (16)$$

Now, the prior distribution of  $\kappa$  is assumed to be Gaussian i.e.  $\kappa \sim \mathcal{N}(a, b^2)$ . The conditional posterior distribution of  $\kappa$ , obtained using Bayes' rule is given below,

$$f_\kappa(\kappa | \boldsymbol{\alpha}, \mathbf{Y}, \sigma) = \frac{1}{\sqrt{2\pi\sigma_\kappa^2}} e^{\frac{-(\kappa - \mu_\kappa)^2}{2\sigma_\kappa^2}} \quad (17)$$

$$\text{where, } \mu_\kappa = \frac{\frac{a}{b^2} + \frac{1}{\sigma^2} \sum_{i=1}^n (\alpha_i - 1)(\mathbf{Y}_i - \alpha_i \mathbf{X}) \cdot \mathbf{1}}{\frac{1}{\sigma^2} \sum_{i=1}^n (\alpha_i - 1)^2 + \frac{1}{b^2}}, \quad \sigma_\kappa^2 = \frac{1}{\frac{1}{\sigma^2} \sum_{i=1}^n (\alpha_i - 1)^2 + \frac{1}{b^2}}$$

$$\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) \quad \text{and} \quad \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$$

The derivation of Equation (17) is given in Appendix .1. Now, it is very clear from the expression of  $\sigma_\kappa^2$  that,  $\sigma_\kappa \leq b$ . The equality holds only when  $\alpha_i = 1$ . So, the variance of the prior distribution of  $\kappa$  is reduced in its posterior distribution.

The prior distribution of  $\alpha_i$  is also assumed to be Gaussian i.e.  $\alpha_i \sim \mathcal{N}(c, d^2)$ . So, the conditional posterior distribution of  $\alpha_i$  is given by,

$$f_\alpha(\alpha_i | \kappa, \mathbf{Y}_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma_{\alpha_i}^2}} e^{\frac{-(\alpha_i - \mu_{\alpha_i})^2}{2\sigma_{\alpha_i}^2}} \quad (18)$$

$$\text{where, } \mu_{\alpha_i} = \frac{\frac{\mathbf{X}^T \mathbf{Y}_i + \kappa(\mathbf{X} + \mathbf{Y}_i) \cdot \mathbf{1} + r\kappa^2}{\sigma^2} + \frac{c}{d^2}}{\frac{1}{d^2} + \frac{\mathbf{X}^T \mathbf{X} + 2\kappa \mathbf{X}^T \mathbf{1} + r\kappa^2}{\sigma^2}}, \quad \sigma_{\alpha_i}^2 = \frac{1}{\frac{1}{d^2} + \frac{\mathbf{X}^T \mathbf{X} + 2\kappa \mathbf{X}^T \mathbf{1} + r\kappa^2}{\sigma^2}}$$

The derivation of Equation (18) is given in Appendix .2. Now,  $\sigma$  is assumed to be uniformly distributed i.e.  $\sigma \sim \mathcal{U}(\theta_1, \theta_2)$ , its conditional posterior distribution is given by,

$$f_{\sigma}(\sigma|\kappa, \boldsymbol{\alpha}, \mathbf{Y}) = \frac{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)} \sigma^{nr}}{\int_{\theta_1}^{\theta_2} \frac{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)} \sigma^{nr}}{d\sigma}} \quad (19)$$

where,  $\sigma \in (\theta_1, \theta_2)$  and  $\mu_i$  is given in Equation (14).

The denominator in the right hand side of Equation (19) can further be solved as the following,

$$\int_{\theta_1}^{\theta_2} \frac{e^{-\frac{\beta}{\sigma^2}}}{\sigma^{nr}} d\sigma = \frac{1}{2} \beta^{\frac{nr-1}{2}} \gamma(1 - \gamma_l - \gamma_u) \quad (20)$$

where,  $\beta$ ,  $\gamma$ ,  $\gamma_l$  and  $\gamma_u$  are given as follows,

$$\beta = \frac{1}{2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i), \quad \gamma = \Gamma\left(\frac{nr-1}{2}\right)$$

$$\gamma_l = \Gamma_{lower}\left(\frac{\beta}{\theta_2^2}, \frac{nr-1}{2}\right), \quad \gamma_u = \Gamma_{upper}\left(\frac{\beta}{\theta_1^2}, \frac{nr-1}{2}\right)$$

The derivation of Equations (19) and (20) is presented in Appendix .3. Now, consider  $\Omega$  to be vector of all the model parameters and  $\Theta$  to be vector of all hyperparameters i.e.

$$\Omega = (\kappa, \sigma, \boldsymbol{\alpha}), \quad \Theta = (a, b, c, d, \theta_1, \theta_2)$$

The joint distribution of  $\boldsymbol{\alpha}$ ,  $\kappa$  and  $\sigma$  is given by,

$$f(\Omega|\Theta, \mathbf{Y}) = f_1(\kappa|\Theta_{\kappa}) f_2(\sigma|\Theta_{\sigma}) \prod_{i=1}^n f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_1(\alpha_i|\Theta_{\alpha}) \quad (21)$$

where,

$$\Theta_{\kappa} = (a, b), \quad \Theta_{\alpha} = (c, d), \quad \Theta_{\sigma} = (\theta_1, \theta_2)$$

$$f_1(x|m, s) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}}, \quad f_2(\sigma|\theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1}$$

Thus, the joint distribution as well as the conditional posterior distributions are obtained. But, the joint posterior distribution of all parameters are required to obtain the Bayes estimates. Even if the joint posterior distribution is evaluated, it will be very difficult to compute marginal posterior distributions from the joint distribution, because it will involve three dimensional integration. One alternative solution can be to simulate the marginal posterior distributions from the condition posterior distributions. Gibbs Sampler is used in this work to simulate the distributions. The following section introduces the Gibbs Sampler and discusses the framework of its application on model parameter estimation. It should be noted that the final Bayes estimate depends on the loss function (22) under consideration, e.g. for square error loss function the final estimate is given by mean of the distribution, whereas, any median of the distribution is the final estimate for absolute error loss function. In this paper, square error loss function is considered for all experiments.

### 3.3. Model Parameter Estimation using Gibbs Sampler

Gibbs sampler (21) is a Markov Chain Monte Carlo (MCMC) algorithm which is used to obtain a sequence of observations from a specified probability distribution. In this approach, previous sample value is used to generate next sample in the sequence, which constructs a Markov Chain. The samples of this Markov Chain converge to the required distribution by construction. For example, suppose a bivariate random variable  $(x, y)$  is considered, and one wishes to compute the marginals,  $p(x)$  and  $p(y)$ . It is far easier to consider a sequence of conditional distributions,  $p(x|y)$  and  $p(y|x)$ , than it is to obtain the marginals by integration of



the joint density  $p(x, y)$ , i.e.

$$p(x) = \int_{\mathbb{D}_y} p(x, y) dy \quad \text{and} \quad p(y) = \int_{\mathbb{D}_x} p(x, y) dx$$

where,  $\mathbb{D}_x$  and  $\mathbb{D}_y$  are the domains of  $x$  and  $y$  respectively. The sampler starts with some initial value  $y_0$  for  $y$  and obtains  $x_1$  by generating a random sample from the conditional distribution  $p(x|y = y_0)$ . The sampler then uses  $x_1$  to generate a new value of  $y_1$ , drawing from the conditional distribution  $p(y|x = x_1)$ . The sampler proceeds as follows

$$x_i \sim p(x|y = y_{i-1}), \quad y_i \sim p(y|x = x_i)$$

Repeating this process  $N$  times, generates a Gibbs sequence of length  $N$ . First  $n(< N)$  terms are rejected to remove the effect of initial guess  $y_0$ . This Gibbs sequence converges to a stationary distribution that is independent of the starting values, and by construction, this stationary distribution is the target distribution which is being simulated. i.e.  $x \sim p(x)$  and  $y \sim p(y)$ . Also, the expectation of any function  $g$  of the random variable  $x$  can be approximated in a similar manner. Using the Law of Large Numbers, expected values of  $x$  and  $g(x)$  can be approximated as follows

$$\frac{1}{N-n} \sum_{i=n+1}^N x_i \xrightarrow{P} E(x) \quad (22)$$

$$\frac{1}{N-n} \sum_{i=n+1}^N g(x_i) \xrightarrow{P} E[g(x)] \quad (23)$$

as  $N \rightarrow \infty$  for some large  $n$ . So, the estimates are obtained by taking average of the generated samples.

### 3.3.1. Algorithm for Model Parameter Estimation using Gibbs Sampler

The steps involved in calculation of the expected value of parameters of the affine model using Gibbs Sampler are enumerated in Algorithm 1.

---

#### Algorithm 2 Parameter Estimation using Gibbs Sampler

---

- 1: **Posterior Distributions** : The posterior distributions of  $\kappa \sim f_\kappa(\kappa|\alpha_1, \dots, \alpha_n, \sigma, \mathbf{Y})$ ,  $\alpha_i \sim f_\alpha(\alpha_i|\kappa, \sigma, \mathbf{Y}_i)$  and  $\sigma \sim f_\sigma(\sigma|\alpha_1, \dots, \alpha_n, \kappa, \mathbf{Y})$  are considered as given in Equations (17), (18) and (19) respectively.
  - 2: **Initial Guess** : Initial guess is made for  $\kappa$  and  $\sigma$  as  $\kappa = \kappa^{(0)}$  and  $\sigma = \sigma^{(0)}$
  - 3: **Sampling of  $\alpha$**  : The  $j$ -th sample of  $\alpha_i$ , i.e.  $\alpha_i^{(j)}$  is generated from its posterior distribution, for  $i = 1, 2, \dots, n$ ,
  - 4:  $\alpha_i^{(j)} \sim f_\alpha(\alpha_i|\kappa^{(j-1)}, \sigma^{(j-1)}, \mathbf{Y}_i)$
  - 5: **Sampling of  $\kappa$**  : The  $j$ -th sample of  $\kappa$ , i.e.  $\kappa^{(j)}$  is generated from its posterior distribution,
  - 6:  $\kappa^{(j)} \sim f_\kappa(\kappa|\alpha_1^{(j)}, \dots, \alpha_n^{(j)}, \sigma^{(j-1)}, \mathbf{Y})$
  - 7: **Sampling of  $\sigma$**  : The  $j$ -th sample of  $\sigma$ , i.e.  $\sigma^{(j)}$  is generated from its posterior distribution,
  - 8:  $\sigma^{(j)} \sim f_\sigma(\sigma|\alpha_1^{(j)}, \dots, \alpha_n^{(j)}, \kappa^{(j)}, \mathbf{Y})$  (23)
  - 9: **Iteration** : The steps 3, 4 and 5 are repeated for  $j = 1, 2, \dots, M$ , where,  $M$  is the number of iterations.
- 

Finally, the expected values of  $\alpha_i$ ,  $\kappa$  and  $\sigma$  are calculated as follows,

$$E(\alpha_i) = \frac{1}{M-m} \sum_{j=m+1}^M \alpha_i^{(j)} \quad (24)$$

$$E(\kappa) = \frac{1}{M-m} \sum_{j=m+1}^M \kappa^{(j)} \quad (25)$$

$$E(\sigma) = \frac{1}{M-m} \sum_{j=m+1}^M \sigma^{(j)} \quad (26)$$

for large  $m$  and  $M$  such that,  $m < M$ . Here,  $m$  is burn-in period.

### 3.3.2. Variation of Model Parameters with respect to Hyperparameters

In order to observe the effects of hyperparameters on the estimated values of model parameters, simulations are performed by varying the value of one hyperparameter while keeping others constant. In simulation, the number of Gibbs runs are 2000 and the burn in period is 1500, i.e. only last 500 values are taken into consideration among 2000 estimated values to get final estimate of  $\kappa$ . The simulation results are shown in Figure 4. The diagrams presented in Figure 4 show the variation in estimated value of  $\kappa$  with hyperparameters  $a$  and  $b$  respectively. The plots indicate that, estimates of  $\kappa$  are highly dependent on both  $a$  and  $b$ . Similar experiments show large dependency of estimates of  $\kappa$ ,  $\alpha$  and  $\sigma$  on other hyperparameters also. From these observations it can be concluded that, hyperparameters need to be estimated first to get better estimate of the model parameters. The estimation of hyperparameters is discussed in the ensuing section.

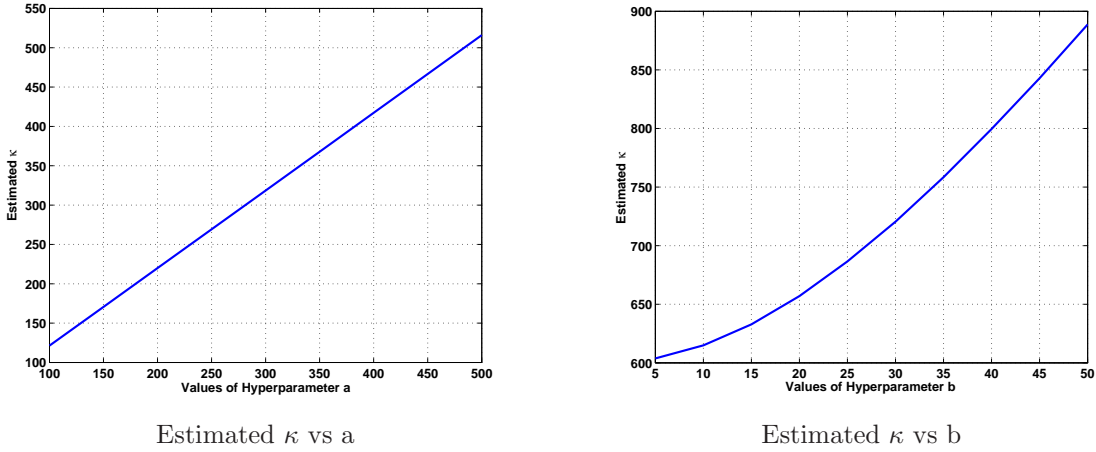


Figure 4: Illustration of variation of estimated values of  $\kappa$  with varying values of hyperparameters

### 3.4. Maximum Likelihood Estimation of Hyperparameters

Hyperparameters of the affine model are estimated using Likelihood Maximization technique. The likelihood function is calculated by multiplying distributions of all the parameters. This function contains both, model parameters and hyperparameters. The model parameters are integrated out to obtain a likelihood function, dependent only on the hyperparameters. The required estimate of the hyperparameters are values for which the likelihood function is maximized.

Under the assumption,  $\kappa \sim \mathcal{N}(a, b^2)$ ,  $\alpha_i \sim \mathcal{N}(c, d^2)$  and  $\sigma \sim \mathcal{U}(\theta_1, \theta_2)$  and  $\epsilon_i \sim \mathcal{N}_r(\mathbf{0}, \sigma^2 \mathbf{I}_{r,r})$ , the likelihood function is given by,

$$L(\Theta|\Omega, \mathbf{Y}) = f_1(\kappa|\Theta_\kappa) f_2(\sigma|\Theta_\sigma) \prod_{i=1}^n f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_1(\alpha_i|\Theta_\alpha) \quad (27)$$

where,  $\Theta, \Omega, \mathbf{Y}, f_1(\kappa|\Theta_\kappa), f_2(\sigma|\Theta_\sigma)$  and  $f_1(\alpha_i|\Theta_\alpha)$  are defined in Section 3. The likelihood function in Equation (27) contains terms of  $a, b, c, d, \theta_1, \theta_2, \kappa, \alpha_1, \dots, \alpha_n$ , and  $\sigma$ . But the required likelihood function

should be a function of hyperparameters and it should not contain model parameters. In order to achieve this, the model parameters  $\kappa, \alpha_1, \dots, \alpha_n$  and  $\sigma$  are integrated out from Equation (27) to obtain the integrated likelihood ( $IL(\Theta)$ ) function as follows,

$$IL(\Theta) = \int_{\theta_1}^{\theta_2} \int_{-\infty}^{\infty} \frac{f(\kappa, \sigma, c, d)}{(\theta_2 - \theta_1)\sqrt{2\pi b^2}} e^{\frac{-(\kappa-a)^2}{2b^2}} d\kappa d\sigma \quad (28)$$

where,  $f(\kappa, \sigma, c, d)$  denotes a part of the likelihood function which is obtained after integrating out all the  $\alpha_i$ 's.

$$f(\kappa, \sigma, c, d) = \prod_{i=1}^n \int_{-\infty}^{\infty} \frac{e^{\frac{-(\alpha_i-c)^2}{2d^2} + \frac{-(\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}}}{\sqrt{2\pi d^2} (2\pi\sigma^2)^{r/2}} d\alpha_i \quad (29)$$

which simplifies to the following expression,

$$f(\kappa, \sigma, c, d) = \frac{A_{\alpha_i}^{-\frac{n}{2}}}{2^{\frac{n(r+1)}{2}} \pi^{\frac{nr}{2}} d^n \sigma^{nr}} e^{\sum_{i=1}^n \left( \frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i} \right)} \quad (30)$$

where,  $A_{\alpha_i} = \frac{1}{2d^2} + \frac{\mathbf{X}^T \mathbf{X} + 2\kappa(\mathbf{X}^T \mathbf{1}) + r\kappa^2}{2\sigma^2}$

$$B_{\alpha_i} = \frac{\mathbf{X}^T \mathbf{Y}_i + \kappa(\mathbf{X} + \mathbf{Y}_i)^T \mathbf{1} + r\kappa^2}{2\sigma^2} + \frac{c}{2d^2}$$

$$C_{\alpha_i} = \frac{\mathbf{Y}_i^T \mathbf{Y}_i + 2\kappa(\mathbf{Y}_i^T \mathbf{1}) + r\kappa^2}{2\sigma^2} + \frac{c^2}{2d^2}$$

Now using Equation (30), Equation (28) can be written as,

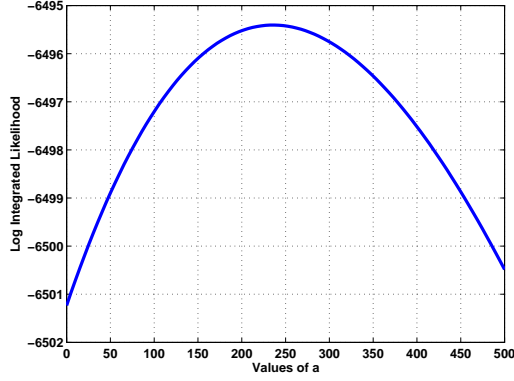
$$IL(\Theta) = \int_{\theta_1}^{\theta_2} \int_{-\infty}^{\infty} \frac{A_{\alpha_i}^{-\frac{n}{2}} e^{\frac{-(\kappa-a)^2}{2b^2} \sum_{i=1}^n \left( \frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i} \right)}}{2^{\frac{n(r+1)+1}{2}} \pi^{\frac{(nr+1)}{2}} d^n \sigma^{nr} (\theta_2 - \theta_1)} d\kappa d\sigma \quad (31)$$

The derivation of Equations (28) through (31) is given in Appendix .4. The values of  $a, b, c, d, \theta_1$  and  $\theta_2$  for which the integrated likelihood displayed in Equation (31) attains maximum value for given  $\mathbf{Y}$ , irrespective of  $\kappa, \alpha_1, \dots, \alpha_n$  and  $\sigma$ , are the final estimates of hyperparameters. Note that, the constant in Equation (31) can be ignored, because it does not affect maximization. So, the integrated likelihood can be written as,

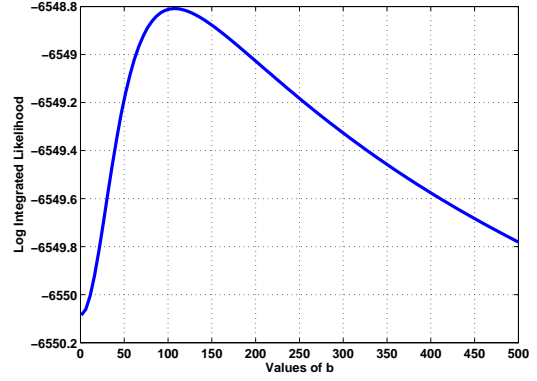
$$IL(\Theta) = \int_{\theta_1}^{\theta_2} \int_{-\infty}^{\infty} \frac{A_{\alpha_i}^{-\frac{n}{2}}}{d^n \sigma^{nr} (\theta_2 - \theta_1)} e^{\frac{-(\kappa-a)^2}{2b^2} \sum_{i=1}^n \left( \frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i} \right)} d\kappa d\sigma \quad (32)$$

Now, instead of calculating  $IL(\Theta)$ ,  $\log(IL(\Theta))$  is considered for optimization, because log is a concave function and it will not affect the maximization. The optimum value of parameters are obtained by maximizing  $\log(IL(\Theta))$ .

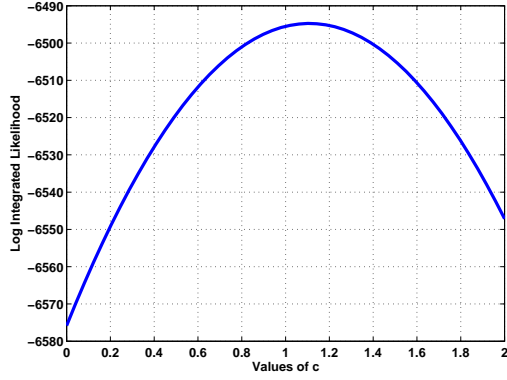
Therefore, the task remains to maximize  $\log(IL(\Theta))$  with respect to  $a, b, c, d, \theta_1$  and  $\theta_2$ . Prior to carrying out the maximization of  $\log(IL(\Theta))$ , nature of the objective function needs to be examined. The variation of  $\log(IL(\Theta))$  against each of the six hyperparameters are shown in Figure 5. A careful observation of these plots indicate that,  $\log(IL(\Theta))$  is uni-modal with respect to all hyperparameters. Hence,  $\log(IL(\Theta))$  can be maximized. A joint maximization is needed over these six variables to obtain the required hyperparameters. Interior-point algorithm (24) has been used to solve this optimization problem. This is a special kind of linear programming algorithm in which the optimal solution is reached by traversing the interior of the feasible region. Here, the feasible region is the set of all possible points of an optimization problem that satisfy the problem's constraints.



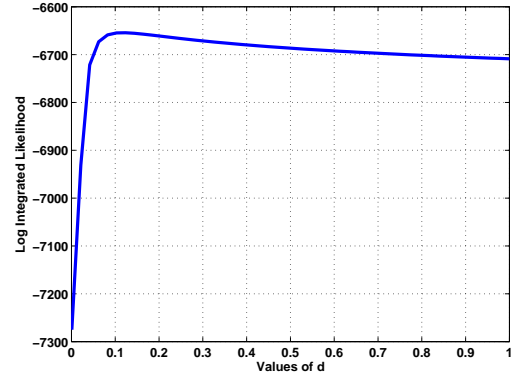
(a)  $\log(IL(\Theta))$  vs  $a$



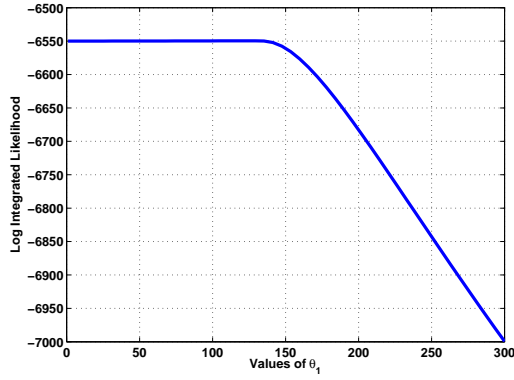
(b)  $\log(IL(\Theta))$  vs  $b$



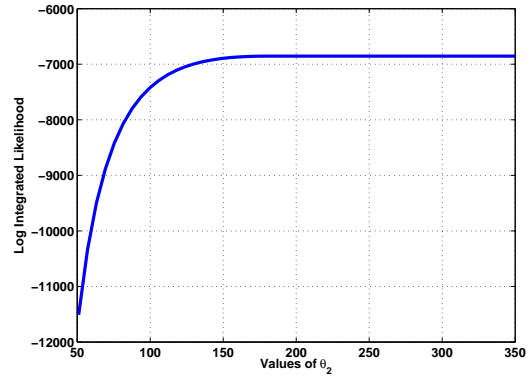
(c)  $\log(IL(\Theta))$  vs  $c$



(d)  $\log(IL(\Theta))$  vs  $d$



(e)  $\log(IL(\Theta))$  vs  $\theta_1$



(f)  $\log(IL(\Theta))$  vs  $\theta_2$

Figure 5: Figures illustrating the variation of Log Integrated Likelihood ( $\log(IL(\Theta))$ ) with respect to  $a$ ,  $b$ ,  $c$ ,  $d$ ,  $\theta_1$  and  $\theta_2$  respectively, varying one at a time keeping others constant

### 3.5. Algorithm to Estimate Model Parameter using Bayesian Estimation Technique

The steps for calculating the affine model parameters using Bayesian Estimation are enumerated in Algorithm 3. This algorithm improves performance of the recognizer by a significant amount. The improvement in recognition accuracy is shown in the following section with the help of two kinds of recognition experiment. Although the number of parameters were kept in check as described in Section 2.1, the computational

cost was higher than the conventional linear VTLN method (3). Thus, both normalization models will find its application depending on the problem under consideration.

---

**Algorithm 3** Estimation of Parameters  $\alpha$  and  $\kappa$  using Bayesian Estimation Method

---

- 1: **Formant Frequency Vectors** : Formant frequencies are extracted (20) from all utterances to construct formant frequency vector for subject speaker ( $\mathbf{X}$ ) and reference speaker ( $\mathbf{Y}$ ).
  - 2: **Integrated likelihood function** : The integrated likelihood function is calculated as shown in Equation (32).
  - 3: **Estimation of Hyperparameters** : The hyperparameters  $a, b, c, d, \theta_1, \theta_2$  are calculated by optimizing Equation (32) using interior point algorithm.
  - 4: **Estimation of  $\alpha$  and  $\kappa$**  : Using the estimates of hyperparameters, the affine model parameters  $\alpha$  and  $\kappa$  are calculated using Algorithm2.
- 

#### 4. Performance Evaluation

In order to evaluate performance of the Bayesian approach for speaker normalization proposed in this paper, two different experiments are conducted namely, vowel recognition experiment and speech recognition experiment. Vowel recognition experiments are conducted to validate the whole set-up of speaker normalization, because it is based on formant frequencies. Afterwards, speech recognition experiments are carried out to demonstrate the scope of the proposed approach for real-life applications.

##### 4.1. Experiments on Vowel Recognition

The vowel recognition experiments are performed using a Mahalanobis distance (25) based vowel recognizer. A brief description is presented on the recognizer as well as the databases being used for vowel recognition and the experimental set-up is discussed. Thereafter the experiments are conducted and recognition performances are shown in terms of recognition accuracy. Note that, the experiments on vowels are solely based on their formant frequencies and not on any other acoustic information from the utterances of vowels.

##### 4.1.1. Formant based Vowel Recognizer

Formant frequency vectors corresponding to each speaker are needed to implement the normalization scheme. The formant frequency vector for a particular speaker is constructed by concatenating formants of all vowels spoken by that speaker. Subsequently, the methods discussed in Sections 2 and 3 are used to compute the normalization parameters for each speaker. The estimation of normalization parameters is followed by its application on formant frequencies of the database to obtain normalized frequencies using Equation (3), as shown in Section 2. First three formant frequencies corresponding to a vowel are considered for recognition. The reference speakers' database contains different vowels spoken by various speakers, which constitutes various instances for each vowel. Finally for testing a vowel utterance, its formant frequencies are extracted and its Mahalanobis distance is computed from each vowel group present in the reference speakers' database. The vowel corresponding to the minimum distance is the recognized vowel.

##### 4.1.2. Vowel Databases

The following two databases are used for vowel recognition experiments.

- *Peterson & Barney Database (PnB)*: There are a total of 76 speakers (33 Males, 28 Females and 15 Children) in PnB database (26). The utterances were recorded using magnetic tape recorder. For the recordings, a list of 10 monosyllabic words were prepared each starting with 'h' and ending with 'd'. Speakers were given a list of words before recording. The order in the lists were randomized, and each speaker was asked to pronounce words using two different lists. Randomizing the list avoids the practice effects of the speakers. This database consisted of utterances of 10 vowels (/aa/, /ae/, /ah/,

/ao/, /eh/, /er/, /ih/, /iy/, /uh/, /uw/), which are uttered twice by each of the speakers. Alternately, the PnB database can be considered to be having 152 speakers (66 Males, 56 Females and 30 Children), with each of them having uttered 10 vowels once.

- *Hillenbrand Database (Hil)*: The Hil database (27) effectively consists of a total of 98 speakers (37 Males, 33 Females and 28 Children). Here each of the speakers have uttered only once, each of the 12 vowels (/ae/, /ah/, /aw/, /eh/, /ei/, /er/, /ih/, /iy/, /oa/, /oo/, /uh/, /uw/). These vowels are extracted from 12 monosyllabic words starting with ‘h’ and ending with ‘d’, which were uttered by those speakers. There were some more speakers in this database, but they are not considered for our experiments as some of the formants corresponding to those speakers are marked zero. The formants were marked zero because the authors were unable to calculate them. The aforementioned databases are available in (28) and (29) respectively.

#### 4.1.3. Experimental Conditions

Each speaker in both databases are characterized by using first three formant frequencies ( $F_1, F_2, F_3$ ) of each vowel. Formant frequency vector corresponding to a speaker is constructed by concatenating the formant frequencies of all vowels spoken by that speaker. Since there are three formant frequencies for each vowel, the dimension of the formant frequency vector corresponding to each speaker will be 30 for PnB database (corresponding to 10 vowels) and 36 for Hil database (corresponding to 12 vowels).

In both databases mentioned above there are three categories of speakers: male, female and child. Thus, the normalization parameters,  $\kappa$  (speaker independent) and  $\alpha$  (speaker dependent) are calculated using all combinations of these three categories, e.g. for male speakers, three different kinds of normalization parameters are obtained by taking male speaker as subject and alternately considering male, female and child speaker as reference. This is followed by normalization of the databases using estimated parameters. Hence, for each kind of speaker there will be three classes of normalized frequencies corresponding to different categories of reference speakers used. In total, there will be 9 different combinations of subject and reference speaker as MM, MF, MC, FM, FF, FC, CM, CF and CC, where M, F, C correspond to Male, Female and Child speaker respectively.

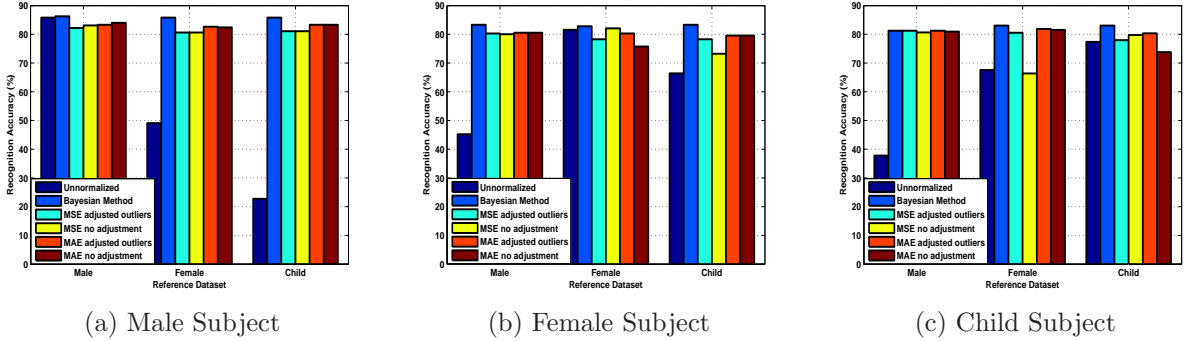


Figure 6: Bar diagrams showing vowel recognition performance on Hil database in terms of Recognition Accuracy for baseline case and using normalization

#### 4.1.4. Vowel Recognition Performance

The recognition experiment on vowels can be divided into two categories, namely

- *Gender Dependent Normalization*: The recognition accuracies are shown using bar diagrams. Figures 6 and 7 show the vowel recognition performances for Hil and PnB databases respectively. In each Figure, there are three bar diagrams corresponding to different kinds of subject speakers. Each bar diagram comprises of three groups corresponding to different categories of reference speakers. The groups in turn contain 6 bars each, among which the first bar corresponds to baseline case and other bars correspond to various normalization methods.

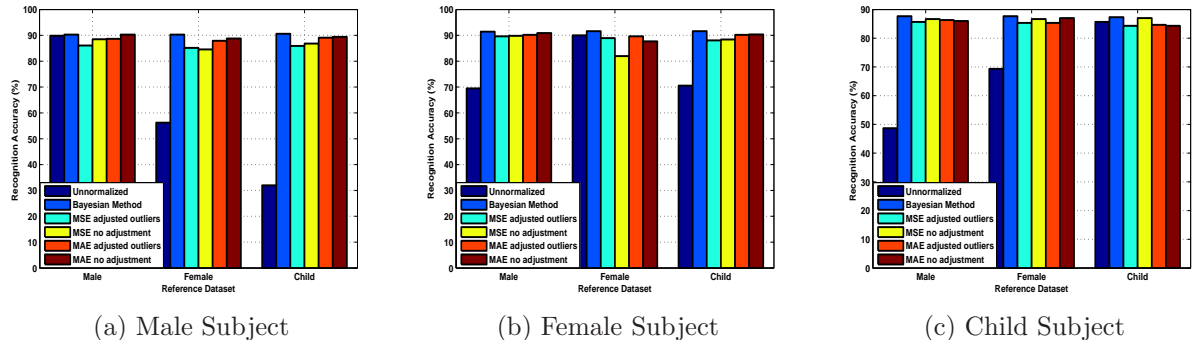


Figure 7: Bar diagrams showing vowel recognition performance on PnB database in terms of Recognition Accuracy for baseline case and using normalization

Table 2: Table illustrating the Vowel Recognition Performance using normalization in terms of Recognition Accuracy (RA) and its improvement over the baseline case

| Normalization Methods      | Hillenbrand Database     |                | Peterson & Barney Database |                |
|----------------------------|--------------------------|----------------|----------------------------|----------------|
|                            | Recognition Accuracy (%) | Improvement(%) | Recognition Accuracy (%)   | Improvement(%) |
| Baseline Case              | 75.2                     | -              | 83.9                       | -              |
| Bayesian Estimation        | 80.1                     | 6.5            | 88.6                       | 5.6            |
| MSE without any adjustment | 75.0                     | -0.3           | 82.4                       | -1.9           |
| MSE with adjusted outliers | 70.2                     | -6.6           | 80.1                       | -4.5           |
| MAE without any adjustment | 78.0                     | 3.7            | 84.3                       | 0.5            |
| MAE with adjusted outliers | 73.3                     | -2.5           | 85.6                       | 2.0            |

- *Gender Independent Normalization:* In this experiment, normalization parameters for a speaker is computed by considering all other speakers present in the database (not of a particular gender) as reference speakers. The recognition accuracy for different experiments are shown in Table 2. The relative improvements in performance over the baseline case (without normalization) while using normalization, are indicated in a separate column. A negative entry in this column demonstrates a degradation in performance.

The experiments discussed in this section signifies that, a greater amount of performance improvement can be achieved for gender dependent normalization compared to gender independent normalization. It also indicates that, normalization parameters estimated using mean absolute error give better performance compared to mean square error.

#### 4.2. Experiments on Speech Recognition

A Hidden Markov Model (HMM) based speech recognizer (30; 31) is used here for speech recognition experiments. A brief description is presented on the speech recognizer as well as the database being used and the experimental set-up is discussed. Thereafter the experiments are carried out and recognition performances are shown in terms of word error rate.

##### 4.2.1. The Recognizer

The standard filter-bank front end introduced by Davis and Mermelstein (32) (which is conventionally used in HMM based speech recognizer) is modified to incorporate normalization method for feature extraction. The normalization scheme is applied on power spectrum of windowed signal during feature extraction. Normalized features are extracted from this modified front-end signal processor. This normalization process changes the bandwidth as well as the frequency bin values of the spectrum. Due to these reasons, normalized spectrum needs to be modified before feature-extraction.

- *Bandwidth Adjustment:* Depending on the value of  $\alpha$ , bandwidth of the spectrum changes. For values of  $\alpha > 1$ , bandwidth increases, whereas  $\alpha < 1$  decreases the bandwidth. This difference in bandwidth is adjusted using piecewise linear warping function. The following function is applied on the warped spectra,

$$G'(f) = \begin{cases} G(f), & 0 \leq f \leq f_0 \\ \frac{f_{max}-G(f_0)}{f_{max}-f_0}(f - f_0) + G(f_0), & f_0 \leq f \leq f_{max} \end{cases} \quad (33)$$

where,  $f_{max}$  is the maximum frequency present in the signal,  $f_0$  is a frequency chosen by the user which falls above the highest significant formant in the speech and  $G(f)$  is the warped spectra, which we get by applying our affine model based normalization with parameters estimated using different methods discussed earlier.

- *Frequency Bin Adjustment:* In general, for the spectrum of  $G(f)$ , the amplitude of the spectrum is available only at specific values of  $f$ . Due to the bandwidth adjustment discussed earlier, the frequency values could no longer be on those specific values. Also, due to the shift in warping function, some of the frequency points might be missing either in the beginning (for positive shift) or in the end (for negative shift) of the spectrum. A simple linear interpolation method is used here to get amplitudes of the spectrum at those values of  $f$  (33; 34) where amplitude of unnormalized spectrum was defined.

The step followed by feature extraction from the acoustic data is training of the HMM models using the features. Subsequently the trained HMM models are used for testing purposes. The training and testing methods are modified as well to incorporate normalization into the recognizer. The process of normalized training and testing is discussed in the following paragraph.

- *Normalized Training:* The training process begins with the computation of formant frequencies from training utterances (20). Then, the formant frequency vector for a speaker is constructed by concatenating formants of all utterances from that speaker. Subsequently, the normalization parameters are estimated using the techniques discussed in Sections 2 and 3. Later normalized MFCC features are extracted from the utterances using normalization parameters which are used to train the HMM model. These steps are summarized in a block diagram shown in Figure 8.
- *Normalized Testing:* In order to incorporate normalization process into testing utterances a two pass approach through the recognizer is adopted. In the first pass, features are extracted from a test utterance to obtain an initial transcript. Utterances corresponding to initial transcript in the training database helps to construct a class of reference formant frequency vectors for the test utterance under consideration. Subsequently, the normalization parameters are computed using the class of reference formant frequency vectors. These parameters are used to extract normalized MFCC features which

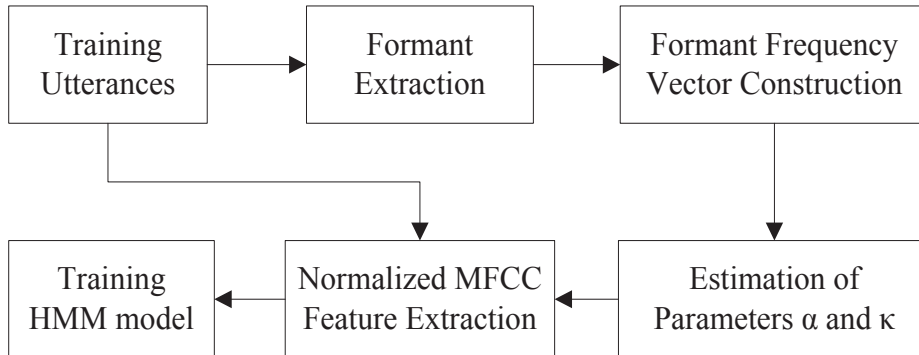


Figure 8: Block Diagram Illustrating Incorporation of Normalization Method for Training the Recognizer



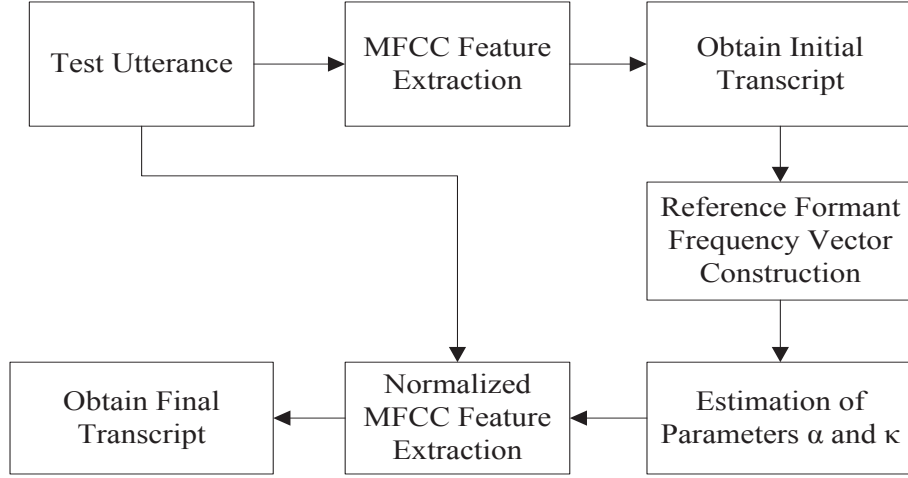


Figure 9: Block Diagram Illustrating Incorporation of Normalization Method for Testing an Utterance using the Recognizer

are again passed through the recognizer to obtain final transcript. Figure 9 summarizes the steps discussed in this module using block diagram.

#### 4.2.2. Hindi Language Database

The database used here, is part of a bigger database, collected for a project, funded by Ministry of Communication & Information Technology, Govt. of India. The goal of the project was to develop a system in which queries regarding prices of some commodities in a particular district can be made through mobile phones using speech modality. A speech database was developed for this purpose, consisting of the names of different commodities and districts. The original database comprises of speech data, collected in six different Indian languages. These languages are spoken in many regional and social dialects, with their own styles. The speech data were collected from about 1000 farmers for each language across different districts to capture the variation in dialects. Farmers were encouraged to do the recordings using their own mobile phones in the environment that they live in. The noise level of data are substantially higher due to the surrounding environment, which can be the field or village country side. The recordings have a sampling rate of 8 kHz and are stored in 16 bit PCM format with a Microsoft .wav header. The challenges posed by this database for the application of speech recognition technology can be summarized as follows,

- Accent and dialect variations.
- High level of background noise.
- Disfluencies and pauses, since the users are naive.
- Issues like poor network coverage, interference, fading etc. and their effect on speech data.

In this work, the experiments are performed on a subset of the above mentioned database, which are recorded in Hindi language. This subset consists of two groups. In one group, there are 107 commodity names (Commodity Database), and the other group consists of 71 district names (District Database). The Commodity database consists of 7431 utterances [ $\sim 38$  hrs], whereas District database has 4783 utterances [ $\sim 24$  hrs].

#### 4.2.3. Experimental Conditions

The databases are divided into two parts, training and testing, as discussed earlier in this section. For Commodity database, the training data consists of 5899 utterances [ $\sim 30.5$  hrs] and testing data consists of 1532 utterances [ $\sim 7.5$  hrs], whereas the training and testing databases for District database consists of 3798

Table 3: Table of specifications for front end signal processing used in the experiments of speech recognition

| Parameter              | Default Value  |
|------------------------|----------------|
| Window Length          | 20ms           |
| Filterbank Type        | Mel Filterbank |
| Number of Mel Filters  | 40             |
| Number of Cepstra      | 13             |
| DFT size               | 512            |
| Lower Filter Frequency | 133.33 Hz      |
| Upper Filter Frequency | 6855.49 Hz     |
| Pre-Emphasis Factor    | 0.97           |

Table 4: Table illustrating the Speech Recognition Performance using normalization in terms of Word Error Rate (WER) and its improvement over the baseline case

| Normalization Methods      | Commodity Database  |                | District Database   |                |
|----------------------------|---------------------|----------------|---------------------|----------------|
|                            | Word Error Rate (%) | Improvement(%) | Word Error Rate (%) | Improvement(%) |
| Baseline Case              | 16.7                | -              | 24.6                | -              |
| Linear VTLN                | 15.5                | 7.2            | 22.5                | 8.5            |
| Bayesian Estimation        | 14.3                | 14.4           | 21.1                | 14.2           |
| MSE without any adjustment | 22.4                | -34.1          | 28.3                | -15.1          |
| MSE with adjusted outliers | 19.8                | -18.6          | 43.2                | -75.6          |
| MAE without any adjustment | 18.6                | -11.4          | 27.5                | -11.8          |
| MAE with adjusted outliers | 17.4                | -4.2           | 38.4                | -56.1          |

[ $\sim 20$  hrs] and 985 [ $\sim 5$  hrs] utterances respectively. The specifications for front end signal processing for feature extraction is given in Table 3. Experiments using context independent phonemes have been carried out. The phonemes in each database are represented using three state left to right HMM. Further, each state of the HMM is modelled using a mixture of 16 Gaussian densities. The experiments are performed using the Sphinx3 toolkit. Note that the conducted experiments are gender independent i.e. the normalization parameters are extracted without gender information of the speaker. This method was adopted to increase usability of the system in real life scenario where such information may not be available.

#### 4.2.4. Speech Recognition Performance

The recognition performance for different experiments is evaluated using Word Error Rate (WER). The results of these experiments are shown in Table 4. First row of this table corresponds to the baseline case, when there is no speaker normalization being used. The following row shows recognition performance using linear normalization model (3). All the following rows display performance using affine model. The parameters of this affine model is estimated using various techniques discussed in Sections 2 and 3. Performance using normalization is compared with baseline case and relative improvements are indicated in a separate column. A negative entry in this column signifies performance degradation. Note that, background noise can significantly degrade the recognition performance. This fact was presented in a work by Hirsch and Pearce, which can be found in (35). In the standard Aurora-2 task for recognising the ten digits and ‘oh’ in American English (i.e. only 11 word vocabulary), they have shown that the performance can vary from 99% to 10% depending on the background noise. In this work also, the speech data in both databases contain large amount of background noise which justifies high WER. The results using normalization indicate that, changing the normalization model from linear to affine, performance can be improved. The parameter estimation method plays an important role in determining the recognition performance using the affine model, and the effect can be easily observed in recognition results presented in Table 4. Among all estimation techniques used, only Bayesian estimation method improves performance of recognizer. This improvement is almost twice than that of using linear model.

## 5. Conclusion

A Bayesian approach to speaker normalization is proposed in this paper. The variation of vocal tract length among different speakers is modelled using an affine model. The parameters of the model are estimated using Error Function Minimization technique and its limitations are discussed. Subsequently, a Bayesian method of parameter estimation is proposed and the framework is described for the problem under consideration. A special type of Markov Chain Monte Carlo method called Gibbs Sampler is used to implement the Bayesian estimation. The need for hyperparameter estimation is also presented. Later, maximum likelihood estimation is used to estimate the hyperparameters corresponding to model parameters.

A Mahalanobis distance based vowel recognizer is introduced and used for the experiments on vowel normalization. First three formant frequencies of a vowel are considered in this kind of recognizer. Experiments are performed for both Gender dependent as well as Gender independent normalization. It is observed that, the improvement in performance in case of gender dependent normalization is higher than that of gender independent normalization. This indicates that, the proposed approach is better suited for cross-gender normalization. The normalization scheme is further used for speech recognition experiments using Hidden Markov Models. Techniques like bandwidth and frequency bin adjustment is used for this purpose. The proposed normalization method requires prior knowledge about the transcript of the utterance under test. Therefore, a two pass approach through the recognizer is proposed to solve this problem. All the experiments discussed in this paper justifies that, Bayesian estimation method gives better performance compared to other methods.

The prior distributions used for all experiments in this work are non-informative, data-driven priors. Currently, methods that use non-Gaussian priors and informative priors are being explored to estimate the speaker normalization parameters. Additionally, the possibility of using higher order speaker normalization models in the proposed Bayesian parameter estimation framework is also being investigated.

*Appendix .1. Derivation of Posterior Distribution of  $\kappa$  given in Equation (17):*

We have  $f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma)$  given in Equation (16) and the prior of  $\kappa$  is Gaussian i.e.  $f_1(\kappa|a, b) = \frac{1}{\sqrt{2\pi}b}e^{-\frac{(\kappa-a)^2}{b^2}}$ . Let  $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2 \dots, \mathbf{Y}_n)$  and  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ . The joint distribution of  $(\mathbf{Y}, \kappa|\boldsymbol{\alpha}, \sigma)$  is given by,

$$\begin{aligned} f_y^{(1)}(\mathbf{Y}, \kappa|\boldsymbol{\alpha}, \sigma) &= \prod_{i=1}^n f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_1(\kappa) \\ &= \frac{1}{(2\pi\sigma^2)^{nr/2}} e^{-\frac{\sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}} \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\kappa-a)^2}{b^2}} \\ &= \frac{1}{(2\pi)^{\frac{nr+1}{2}} b \sigma^{nr}} e^{-(\kappa^2 A_\kappa - 2\kappa B_\kappa + C_\kappa)} \\ &= \frac{1}{(2\pi)^{\frac{nr+1}{2}} b \sigma^{nr}} e^{-A_\kappa (\kappa - \frac{B_\kappa}{A_\kappa})^2 + (\frac{B_\kappa^2}{A_\kappa} - C_\kappa)} \end{aligned} \quad (.1)$$

where,  $A_\kappa = \frac{1}{\frac{r}{2\sigma^2} \sum_{i=1}^n (\alpha_i - 1)^2 + \frac{1}{2b^2}}$

$$B_\kappa = \frac{a}{2b^2} + \frac{1}{2\sigma^2} \sum_{i=1}^n (\alpha_i - 1) \{(\mathbf{Y}_i - \alpha_i \mathbf{X})^T \mathbf{1}\}$$

$$C_\kappa = \frac{a^2}{2b^2} + \frac{1}{2\sigma^2} \sum_{i=1}^n \{(\mathbf{Y}_i - \alpha_i \mathbf{X})^T (\mathbf{Y}_i - \alpha_i \mathbf{X})\}$$

Now the joint distribution of  $(\mathbf{Y}|\boldsymbol{\alpha}, \sigma)$  can be obtained by integrating Equation (.1) over  $\kappa$  as follows,

$$\begin{aligned}
f_y^{(2)}(\mathbf{Y}|\boldsymbol{\alpha}, \sigma) &= \int_{-\infty}^{\infty} f_y^{(1)}(\mathbf{Y}, \kappa|\boldsymbol{\alpha}, \sigma) d\kappa \\
&= \frac{1}{(2\pi)^{\frac{nr+1}{2}} b\sigma^{nr}} \int_{-\infty}^{\infty} e^{-A_\kappa(\kappa - \frac{B_\kappa}{A_\kappa})^2 + (\frac{B_\kappa^2}{A_\kappa} - C_\kappa)} d\kappa \\
&= \frac{1}{(2\pi)^{\frac{nr+1}{2}} b\sigma^{nr}} \left( \sqrt{\frac{\pi}{A_\kappa}} \right) e^{(\frac{B_\kappa^2}{A_\kappa} - C_\kappa)}
\end{aligned} \tag{.2}$$

Finally, posterior distribution of  $\kappa$  can be obtained by dividing Equation (.1) by Equation (.2)

$$\begin{aligned}
f_\kappa(\kappa|\mathbf{Y}, \boldsymbol{\alpha}, \sigma) &= \frac{f_y^{(1)}(\mathbf{Y}, \kappa|\boldsymbol{\alpha}, \sigma)}{f_y^{(2)}(\mathbf{Y}|\boldsymbol{\alpha}, \sigma)} \\
&= \left( \sqrt{\frac{A_\kappa}{\pi}} \right) e^{-A_\kappa(\kappa - \frac{B_\kappa}{A_\kappa})^2} \\
&= \frac{1}{\sqrt{2\pi\sigma_{post}^2}} e^{-\frac{(\kappa - \mu_{post})^2}{2\sigma_{post}^2}}
\end{aligned} \tag{.3}$$

where,  $\sigma_\kappa^2 = \frac{1}{2A_\kappa}$  and  $\mu_\kappa = \frac{B_\kappa}{A_\kappa}$

*Appendix .2. Derivation of Posterior Distribution of  $\alpha_i$  given in Equation (18):*

The prior distribution of  $\alpha_i$  is assumed to be Gaussian i.e.  $f_1(\alpha_i|c, d) = \frac{1}{\sqrt{2\pi}d} e^{-\frac{(\alpha_i - c)^2}{d^2}}$ . So, the joint distribution of  $\mathbf{Y}_i$  and  $\alpha_i$  can be obtained by multiplying  $f_1(\alpha_i)$  with  $f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma)$  as follows,

$$\begin{aligned}
f_\alpha^{(1)}(\mathbf{Y}_i, \alpha_i|\kappa, \sigma) &= f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_1(\alpha_i) \\
&= \frac{1}{(2\pi\sigma^2)^{r/2}} e^{-\frac{(\mathbf{Y}_i - \mu_i)^T(\mathbf{Y}_i - \mu_i)}{2\sigma^2}} \frac{1}{\sqrt{2\pi}d} e^{-\frac{(\alpha_i - c)^2}{d^2}} \\
&= \frac{1}{(2\pi)^{\frac{r+1}{2}} d\sigma^r} e^{-(A_{\alpha_i}\alpha_i^2 - 2B_{\alpha_i}\alpha_i + C_{\alpha_i})} \\
&= \frac{1}{(2\pi)^{\frac{r+1}{2}} d\sigma^r} e^{-A_{\alpha_i}(\alpha_i - \frac{B_{\alpha_i}}{A_{\alpha_i}})^2 + (\frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i})}
\end{aligned} \tag{.4}$$

where,  $A_{\alpha_i} = \frac{1}{2d^2} + \frac{\mathbf{X}^T \mathbf{X} + 2\kappa(\mathbf{X}^T \mathbf{1}) + r\kappa^2}{2\sigma^2}$

$$B_{\alpha_i} = \frac{\mathbf{X}^T \mathbf{Y}_i + \kappa(\mathbf{X} + \mathbf{Y}_i)^T \mathbf{1} + r\kappa^2}{2\sigma^2} + \frac{c}{2d^2}$$

$$C_{\alpha_i} = \frac{\mathbf{Y}_i^T \mathbf{Y}_i + 2\kappa(\mathbf{Y}_i^T \mathbf{1}) + r\kappa^2}{2\sigma^2} + \frac{c^2}{2d^2}$$

Now the distribution of  $(\mathbf{Y}_i|\kappa, \sigma)$  can be obtained by integrating Equation (.4) over  $\alpha_i$  as follows,

$$\begin{aligned}
f_\alpha^{(2)}(\mathbf{Y}_i|\kappa, \sigma) &= \int_{-\infty}^{\infty} f_\alpha^{(1)}(\mathbf{Y}_i, \alpha_i|\kappa, \sigma) d\alpha_i \\
&= \frac{1}{(2\pi)^{\frac{r+1}{2}} d\sigma^r} \int_{-\infty}^{\infty} e^{-A_{\alpha_i}(\alpha_i - \frac{B_{\alpha_i}}{A_{\alpha_i}})^2 + (\frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i})} d\alpha_i \\
&= \frac{1}{(2\pi)^{\frac{r+1}{2}} d\sigma^r} \left( \sqrt{\frac{\pi}{A_{\alpha_i}}} \right) e^{(\frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i})}
\end{aligned} \tag{.5}$$

Finally, posterior distribution of  $\alpha_i$  can be obtained by dividing Equation (.4) by Equation (.5)

$$\begin{aligned}
f_{\alpha}(\alpha_i|\kappa, \sigma, \mathbf{Y}_i) &= \frac{f_{\alpha}^{(1)}(\mathbf{Y}_i, \alpha_i|\kappa, \sigma)}{f_{\alpha}^{(2)}(\mathbf{Y}_i|\kappa, \sigma)} \\
&= \left( \sqrt{\frac{A_{\alpha_i}}{\pi}} \right) e^{-A_{\alpha_i}(\kappa - \frac{B_{\alpha_i}}{A_{\alpha_i}})^2} \\
&= \frac{1}{\sqrt{2\pi\sigma_{post}^2}} e^{-\frac{(\alpha_i - \mu_{post})^2}{2\sigma_{post}^2}}
\end{aligned} \tag{.6}$$

where,  $\sigma_{\alpha_i}^2 = \frac{1}{2A_{\alpha_i}}$  and  $\mu_{\alpha_i} = \frac{B_{\alpha_i}}{A_{\alpha_i}}$

*Appendix .3. Derivation of Posterior Distribution of  $\sigma$  given in Equation (19):*

The prior distribution of  $\sigma$  is assumed to be uniform i.e.  $f_2(\sigma|\theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1}$ . So, the joint distribution of  $\mathbf{Y}_i$  and  $\sigma$  can be obtained by multiplying  $f_2(\sigma|\theta_1, \theta_2)$  with  $f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma)$  as follows,

$$\begin{aligned}
f_{\sigma}^{(1)}(\mathbf{Y}, \sigma|\alpha, \kappa) &= \prod_{i=1}^n f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_2(\sigma) \\
&= \frac{1}{(2\pi\sigma^2)^{nr/2}} e^{-\frac{\sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}} \frac{1}{\theta_2 - \theta_1} \\
&= \frac{1}{(2\pi)^{\frac{nr}{2}} (\theta_2 - \theta_1) \sigma^{nr}} e^{-\frac{\sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}}
\end{aligned} \tag{.7}$$

Now the distribution of  $(\mathbf{Y}|\alpha, \kappa)$  can be obtained by integrating Equation (.7) over  $\sigma$  as follows,

$$\begin{aligned}
f_{\sigma}^{(2)}(\mathbf{Y}|\alpha, \kappa) &= \int_{-\infty}^{\infty} f_{\sigma}^{(1)}(\mathbf{Y}, \sigma|\alpha, \kappa) d\sigma \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{nr}{2}} (\theta_2 - \theta_1) \sigma^{nr}} e^{-\frac{\sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}} d\sigma
\end{aligned} \tag{.8}$$

Finally, posterior distribution of  $\alpha_i$  can be obtained by dividing Equation (.7) by Equation (.8)

$$\begin{aligned}
f_{\sigma}(\sigma|\kappa, \alpha, \mathbf{Y}) &= \frac{f_{\sigma}^{(1)}(\mathbf{Y}, \sigma|\alpha, \kappa)}{f_{\sigma}^{(2)}(\mathbf{Y}|\alpha, \kappa)} \\
&= \frac{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}}{\sigma^{nr}} \\
&= \frac{\int_{\theta_1}^{\theta_2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)} \sigma^{nr} d\sigma}{\int_{\theta_1}^{\theta_2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)} \sigma^{nr} d\sigma}
\end{aligned} \tag{.9}$$

Assuming  $\beta = \frac{1}{2} \sum_{i=1}^n (\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)$  the denominator in the above Equation can be written as,

$$\begin{aligned}
\int_{\theta_1}^{\theta_2} \frac{e^{-\frac{\beta}{\sigma^2}}}{\sigma^{nr}} d\sigma &= \frac{1}{2} \beta^{\frac{1-nr}{2}} \int_{\frac{\beta}{\theta_1^2}}^{\frac{\beta}{\theta_2^2}} z^{\frac{nr-1}{2}-1} e^{-z} dz \\
&= \frac{1}{2} \beta^{\frac{1-nr}{2}} (\gamma_l - \gamma_u)
\end{aligned} \tag{.10}$$

where,  $\gamma = \Gamma\left(\frac{nr-1}{2}\right)$ ,  $\gamma_l = \Gamma_{lower}\left(\frac{\beta}{\theta_2^2}, \frac{nr-1}{2}\right)$  and  $\gamma_u = \Gamma_{upper}\left(\frac{\beta}{\theta_1^2}, \frac{nr-1}{2}\right)$

*Appendix .4. Derivation of Integrated Likelihood Function given in Equation (28):*

The likelihood function of  $\Theta$  given in Equation (27) can be written as follows,

$$\begin{aligned}
L(\Theta|\Omega, \mathbf{Y}) &= f_1(\kappa|\Theta_\kappa) f_2(\sigma|\Theta_\sigma) \prod_{i=1}^n f_y(\mathbf{Y}_i|\kappa, \alpha_i, \sigma) f_1(\alpha_i|\Theta_\alpha) \\
&= \left( \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\kappa-a)^2}{b^2}} \frac{1}{\theta_2 - \theta_1} \right) \left( \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{r/2}} e^{-\frac{(\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2}} \frac{1}{\sqrt{2\pi}d} e^{-\frac{(\alpha_i - c)^2}{d^2}} \right)
\end{aligned} \tag{.11}$$

Equation (.11) is integrated over  $\alpha$ ,  $\kappa$  and  $\sigma$  to obtain an integrated likelihood function solely dependent on hyperparameters as given below,

$$\begin{aligned}
IL(\Theta) &= \int_{\theta_1}^{\theta_2} \frac{1}{\theta_2 - \theta_1} \left( \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\kappa-a)^2}{b^2}} \left( \prod_{i=1}^n \int_{-\infty}^{\infty} \frac{e^{-\frac{(\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2} - \frac{(\alpha_i - c)^2}{2d^2}}}{(2\pi\sigma^2)^{r/2} \sqrt{2\pi}d} d\alpha_i \right) d\kappa \right) d\sigma \\
&= \int_{\theta_1}^{\theta_2} \frac{1}{\theta_2 - \theta_1} \left( \int_{-\infty}^{\infty} \frac{f(\kappa, \sigma, c, d)}{\sqrt{2\pi}b} e^{-\frac{(\kappa-a)^2}{b^2}} d\kappa \right) d\sigma
\end{aligned} \tag{.12}$$

The function,  $f(\kappa, \sigma, c, d)$  in Equation (.12) is calculated as follows,

$$\begin{aligned}
f(\kappa, \sigma, c, d) &= \prod_{i=1}^n \int_{-\infty}^{\infty} \frac{e^{-\frac{(\mathbf{Y}_i - \mu_i)^T (\mathbf{Y}_i - \mu_i)}{2\sigma^2} - \frac{(\alpha_i - c)^2}{2d^2}}}{(2\pi\sigma^2)^{r/2} \sqrt{2\pi}d} d\alpha_i \\
&= \prod_{i=1}^n \frac{1}{(2\pi)^{\frac{r+1}{2}} d\sigma^r} \left( \sqrt{\frac{\pi}{A_{\alpha_i}}} \right) e^{\left( \frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i} \right)} \\
&= \frac{A_{\alpha_i}^{-\frac{n}{2}}}{2^{\frac{n(r+1)}{2}} \pi^{\frac{nr}{2}} d^n \sigma^{nr}} e^{\sum_{i=1}^n \left( \frac{B_{\alpha_i}^2}{A_{\alpha_i}} - C_{\alpha_i} \right)}
\end{aligned} \tag{.13}$$

- [1] D. O'Shaughnessy, Interacting with computers by voice: automatic speech recognition and synthesis, Proceedings of the IEEE 91 (9) (2003) 1272–1305.
- [2] J. Cohen, T. Kamm, A. Andreou, An experiment in systematic speaker variability, in: Final Day Review., DoD Speech Workshop on Robust Speech Recognition, Baltimore, 1994.
- [3] L. Lee, R. C. Rose, Speaker normalization using efficient frequency warping procedures, in: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol. 1, IEEE, 1996, pp. 353–356.
- [4] S. Umesh, S. Bharath Kumar, M. Vinay, R. Sharma, R. Sinha, A simple approach to non-uniform vowel normalization, in: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, Vol. 1, IEEE, 2002, pp. 1–517.
- [5] P. E. Nordstrom, B. Lindblom, A normalization procedure for vowel formant data, Int. Cong. Phonetic Sci., Leeds, England.
- [6] G. Fant, Non-uniform vowel normalization, Speech Trans. Lab. Q. Prog. Stat. Rep (1975) 2–3.
- [7] J. D. Miller, Auditory-perceptual interpretation of the vowel, The journal of the Acoustical society of America 85 (1989) 2114.
- [8] N. Flynn, Comparing vowel formant normalisation procedures, York Pap. Linguist., Ser 2 (11) (2011) 1–28.
- [9] L. Lee, R. Rose, A frequency warping approach to speaker normalization, Speech and audio processing, IEEE transactions on 6 (1) (1998) 49–60.
- [10] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, Speaker normalization on conversational telephone speech, in: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol. 1, IEEE, 1996, pp. 339–341.
- [11] R. Sinha, S. Umesh, Non-uniform scaling based speaker normalization, in: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, Vol. 1, IEEE, 2002, pp. 1–589.
- [12] S. Umesh, L. Cohen, D. Nelson, Fitting the mel scale, in: Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, Vol. 1, IEEE, 1999, pp. 217–220.
- [13] E. Eide, H. Gish, A parametric approach to vocal tract length normalization, in: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol. 1, IEEE, 1996, pp. 346–348.
- [14] S. S. Stevens, J. Volkmann, The relation of pitch of frequency: A revised scale, Am. J. Psychol.
- [15] A. Acero, R. M. Stern, Robust speech recognition by normalization of the acoustic space, in: Acoustics, Speech, and

- Signal Processing, 1991. ICASSP-91., 1991 International Conference on, IEEE, 1991, pp. 893–896.
- [16] S. Cox, Speaker normalization in the mfcc domain, in: Proc. Int. Conf. on Spoken Language Processing, Vol. 2, 2000, pp. 853–856.
  - [17] D. Sanand, D. D. Kumar, S. Umesh, Linear transformation approach to vtln using dynamic frequency warping, in: Eighth Annual Conference of the International Speech Communication Association, 2007.
  - [18] S. B. Kumar, S. Umesh, Nonuniform speaker normalization using affine transformation, The Journal of the Acoustical Society of America 124 (2008) 1727.
  - [19] J. A. Nelder, R. Mead, A simplex method for function minimization, Computer journal 7 (4) (1965) 308–313.
  - [20] R. C. Snell, F. Milinazzo, Formant location from lpc analysis data, Speech and Audio Processing, IEEE Transactions on 1 (2) (1993) 129–134.
  - [21] G. Casella, E. I. George, Explaining the gibbs sampler, The American Statistician 46 (3) (1992) 167–174.
  - [22] R. V. Hogg, A. T. Craig, Introduction to mathematical statistics. 1978.
  - [23] S. G. W. Paul Damien, Sampling truncated normal, beta, and gamma densities, Journal of Computational and Graphical Statistics 10 (2) (2001) 206–215. doi:10.1198/10618600152627906.
  - [24] N. Karmarkar, A new polynomial-time algorithm for linear programming, in: Proceedings of the sixteenth annual ACM symposium on Theory of computing, ACM, 1984, pp. 302–311.
  - [25] P. C. Mahalanobis, On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India (1936) 49–55.
  - [26] G. E. Peterson, H. L. Barney, Control methods used in a study of the vowels, The Journal of the Acoustical Society of America 24 (1952) 175.
  - [27] J. Hillenbrand, L. A. Getty, M. J. Clark, K. Wheeler, Acoustic characteristics of american english vowels, The Journal of the Acoustical society of America 97 (1995) 3099.
  - [28] Peterson and barney database, <http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/speech/database/pb/0.html>.
  - [29] Hillenbrand database, <http://homepages.wmich.edu/~hillenbr/voweldata.html>.
  - [30] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.
  - [31] J. Mariani, Recent advances in speech processing, in: Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on, IEEE, 1989, pp. 429–440.
  - [32] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, Acoustics, Speech and Signal Processing, IEEE Transactions on 28 (4) (1980) 357–366.
  - [33] P. Zhan, M. Westphal, Speaker normalization based on frequency warping, in: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, Vol. 2, IEEE, 1997, pp. 1039–1042.
  - [34] P. Zhan, A. Waibel, Vocal tract length normalization for large vocabulary continuous speech recognition, Tech. rep., DTIC Document (1997).
  - [35] H. G. Hirsch, D. Pearce, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), 2000.